10.12871/CLICIT2014168

Il *Perugia Corpus*: una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione

Stefania Spina
Dipartimento di Scienze Umane e Sociali, Università per Stranieri di Perugia
stefania.spina@unistrapg.it

Abstract

Italiano Il *Perugia Corpus* (PEC) è un corpus dell'italiano contemporaneo scritto e parlato, che comprende oltre 26 milioni di parole. L'obiettivo che ha guidato la sua costituzione è quello di ovviare alla mancanza di un corpus di riferimento dell'italiano. In questo articolo vengono descritti i criteri alla base della sua composizione, la sua strutturazione in 10 sezioni e sottosezioni e la sua annotazione multilivello, con la relativa valutazione.

English *The Perugia Corpus (PEC) is a corpus of contemporary written and spoken Italian of more than 26 million words. Its aim is to fill the gap of the lack of an Italian reference corpus. This paper describes its composition and organization in 10 sections and sub-sections, and its multilevel annotation and evaluation.*

1 Introduzione

Il Perugia Corpus (PEC) è un corpus di riferimento dell'italiano contemporaneo, scritto e parlato[1]; è composto da oltre 26 milioni di parole, distribuite in 10 differenti sezioni, corrispondenti ad altrettanti generi testuali, e dotato di una annotazione multilivello. Il PEC intende ovviare alla mancanza di un corpus di riferimento (scritto e parlato), di cui hanno finora sofferto gli studi sull'italiano. Per la sua natura di risorsa di riferimento (EAGLES, 1996), il PEC è progettato per fornire informazioni linguistiche il più possibile generali sull'italiano e le sue principali varietà scritte e parlate.

La filosofia che ha guidato la composizione del PEC è dunque radicalmente diversa da quella che è alla base di alcuni web corpora (Baroni e Bernardini, 2006; Kilgarriff e Grefenstette, 2003) dell'italiano di ultima generazione come *Paisà* (Lyding et al., 2014), *itWac* (Baroni e Kilgarriff, 2006) o *itTenTen* (Jakubíček et al., 2013), ma anche da quella di corpora meno recenti come *Repubblica* (Baroni et al., 2004) e CORIS/CODIS (Rossini Favretti et al., 2002): la scelta è stata infatti quella di privilegiare la differenziazione dei generi testuali, includendo anche il parlato, a scapito delle dimensioni del corpus. Inoltre, si è puntato sulla riutilizzazione di risorse già esistenti e disponibili (Zampolli, 1991), ma a volte disperse e di difficile consultazione; ad esse sono stati aggiunti dati nuovi, raccolti col duplice scopo di riempire vuoti in cui non erano disponibili dati per l'italiano, ed aggiornare risorse esistenti, ma ormai datate. Il PEC può dunque essere considerato un corpus di riferimento "low cost", di dimensioni contenute ma con una buona rappresentatività delle diverse varietà scritte e parlate dell'italiano. Le dimensioni contenute del PEC presentano inoltre due vantaggi: permettono di gestire, in fase di interrogazione, quantità di risultati più maneggevoli (Hundt e Leech, 2012), e consentono di ottenere una buona accuratezza nell'annotazione (vedi par. 3.2).

2 Composizione del corpus

Il PEC è suddiviso in 10 sezioni, a loro volta articolate in sottosezioni; complessivamente, i testi inseriti nel corpus sono 41.401, con una lunghezza media di 12.500 tokens per testo. In linea con quanto avviene per corpora di riferimento di altre lingue, anche di dimensioni maggiori, come il *British National Corpus* (Burnard, 2007), lo scritto copre l'85% del totale del PEC, ed il parlato il restante 15%. La tab.1 presenta un quadro riassuntivo del corpus, con i dati relativi alle 10 sezioni; nei paragrafi che seguono, sarà invece descritta, per ogni sezione, la sua composizione interna.

[1] Il PEC è stato realizzato all'Università per Stranieri di Perugia tra il 2011 e il 2012.

sezione	n. testi	tokens	media tokens	% totale	types	TTR	frasi	tokens x frase
SCRITTO								
letteratura	60	3.545.459	59.091	13,38	103.141	54,78	229.361	15,46
saggi	79	2.354.996	29.810	8,89	97.795	63,73	102.130	23,06
stampa	8.232	5.772.040	701	21,78	147.707	61,48	225.827	25,56
accademico	240	1.113.590	4.640	4,20	54.658	51,80	32.736	34,02
scuola	4.054	1.257.842	310	4,75	46.981	41,89	51.208	24,56
amministrazione	119	1.160.334	9.751	4,38	28.562	26,52	31.950	36,32
web	27.383	7.359.460	269	27,78	225.190	83,01	295.041	29,94
TOT. SCRITTO	*40.167*	*22.563.721*		*85,16*	*704.034*		*969.059*	
PARLATO								
tv	127	1.147.151	9.033	4,33	50.643	47,28	73.950	15,51
film	66	626.487	9.492	2,36	31.967	40,39	99.858	6,27
parlato	1.041	2.158.522	2.074	8,15	67.987	46,28	80.354	26,86
TOT. PARLATO	*1.234*	*3.932.160*		*14,84*	*150.597*		*254.162*	
TOTALE	**41.401**	**26.495.881**	**12.517**		**854.631**		**1.223.221**	

Tabella 1 - La composizione delle 10 sezioni del PEC; la type-token ratio (TTR) è calcolata usando l'indice di Guiraud (Guiraud, 1954), per ovviare alla non omogeneità nel numero dei tokens.

2.1 Letteratura

La sezione dedicata alla letteratura comprende campioni estratti da 60 romanzi contemporanei, pubblicati tra il 1990 e il 2012 da 45 autori italiani diversi.

2.2 Saggistica

La saggistica comprende campioni estratti da 79 saggi di argomento diverso, ma riconducibili a quattro aree tematiche (attualità, biografia, politica e tempo libero). Tutti i saggi sono stati pubblicati da autori italiani dal 1990 al 2010.

2.3 Stampa

Gli 8.232 testi della sezione della stampa sono suddivisi tra articoli di quotidiani (79%) e di settimanali (21%): sono infatti tratti dal *Corriere della Sera* e da *Il Sole 24 ore* del 2012, e da *L'Espresso* del 2011 e del 2012. La tab. 2 riporta l'ulteriore suddivisione degli articoli dei quotidiani in 9 sottocategorie, con il rispettivo numero di tokens.

argomento	tokens	% totale
editoriale	436.570	7,6
politica	1.023.021	17,7
economia	565.641	9,8
cronaca	1.555.654	27,0
esteri	681.769	11,8
cultura	667.181	11,6
sport	424.416	7,4
lettere	120.178	2,1
spettacolo	297.610	5,2

Tabella 2 - Tipologie di articoli di quotidiani

2.4 Scritto accademico

In questa sezione è stato incorporato e riutilizzato integralmente, con alcune integrazioni, il *Corpus di Italiano Accademico* (Spina, 2010). In essa sono incluse quattro sottosezioni (tesi di laurea, dispense, manuali e articoli scientifici), a loro volta ripartite fra tre macroaree tematiche (umanistica, giuridico-economica e scientifica). La tab. 3 riporta i dati delle varie sottosezioni.

	tesi	dispense	manuali	articoli	TOT.
umanistica	55.311	170.501	36.077	114.581	*376.470*
giur-eco	58.087	176.206	64.020	75.814	*374.127*
scientifica	54.460	203.197	34.464	70.872	*362.993*
TOT.	*167.858*	*549.904*	*134.561*	*261.267*	*1.113590*

Tabella 3 - Sottosezioni dello scritto accademico

2.5 Scritto scolastico

La sezione è costituita da 4.054 temi svolti da studenti delle scuole medie e superiori tra il 2010 e il 2011, su 21 argomenti diversi; i temi sono stati estratti in modo automatico dal sito di *Repubblica Scuola*. Le due sottosezioni in cui la sezione è articolata sono i 2.431 temi della scuola media (652.749 tokens) e i 1.623 della scuola superiore (605.093 tokens).

2.6 Scritto amministrativo

La sezione amministrativa è composta per il 75% da testi di leggi (europee, statali, regionali), e per il restante 25% da regolamenti e documenti amministrativi più brevi.

2.7 Web

I testi scritti estratti dal web rappresentano la sezione più ampia del PEC, a sua volta suddivisa in testi di interazione e testi di riferimento, come

descritto nella tab. 4. Per quanto riguarda i blog, sono stati estratti i soli testi dei post, senza i commenti, da una cinquantina di blog di genere personale, giornalistico o aziendale. I testi di Wikipedia sono stati prelevati nel gennaio 2012 dalla versione italiana integrale, e selezionati in modo casuale. La sottosezione dei social network comprende 24.424 post tratti da profili di Facebook e di Twitter delle tre tipologie personale, politico e aziendale.

	tokens	% totale
INTERAZIONE		
blog	2.812.439	38,22
forum	171.111	2,33
chat	119.279	1,62
social network	603.630	8,20
TOT. INTERATTIVI	*3.706.459*	*50,36*
RIFERIMENTO		
Wikipedia	3.653.001	49,64
TOT. RIFERIMENTO	*3.653.001*	*49,64*

Tabella 4 - Sottosezioni della sezione web

2.8 Parlato

Per la sezione del parlato si è fatto ampio ricorso a corpora già esistenti e disponibili per uso accademico: il PEC contiene infatti i seguenti corpora o materiali testuali già trascritti, pari circa a 450.000 tokens:

- i testi del *LIP* (De Mauro et al., 1993), nella versione resa disponibile dal sito *Badip* (http://badip.uni-graz.at/it/);
- la sezione italiana del corpus *Saccodeyl*, un progetto *Minerva* sulla lingua dei giovani europei (Pérez-Paredes e Alcaraz-Calero, 2007);
- alcune trascrizioni del corpus CLIPS (Albano Leoni, 2007), tratte dalle sezioni elicitate attraverso map task e test delle differenze.

Questi dati già esistenti sono stati (ri)annotati secondo i criteri previsti dal PEC ed aggiunti al resto dei testi, raccolti ex novo.

La bipartizione principale della sezione del parlato è quella tra parlato dialogico (1.020.264 tokens) e parlato monologico (1.138.258 tokens); ad un livello successivo, il parlato dialogico, sulla base di una distinzione fondamentale derivata dall'analisi della conversazione (Drew e Heritage, 1992), è stato suddiviso in dialogo tra pari (faccia a faccia o telefonico) e dialogo istituzionale (in vari contesti, come quello scolastico-accademico, processuale, medico ecc.). Il parlato monologico, invece, è suddiviso nelle 7 sottosezioni descritte nella tab. 5.

	tokens	% tot.
DIALOGICO		
a. tra pari	*471.097*	*21,82*
- faccia a faccia	187.454	8,68
- telefonico	283.643	13,14
b. istituzionale (lezioni, processi...)	*549.167*	*25,44*
TOT. DIALOGICO	*1.020.264*	*47,27*
MONOLOGICO		
conferenze	168.051	7,79
lezioni	156.128	7,23
processi	174.728	8,09
istituzioni	158.967	7,36
politica	158.346	7,34
religione	168.917	7,83
testi di canzoni[2]	153.121	7,09
TOT. MONOLOGICO	*1.138.258*	*52,73*

Tabella 5 - Sottosezioni del parlato

2.9 Televisione

I dati televisivi inclusi nel PEC derivano dal *Corpus di Italiano Televisivo* (Spina, 2005), in una versione riorganizzata e ampliata. Le 127 trasmissioni comprese nel PEC appartengono ai due macrogeneri "informazione" e "intrattenimento", e sono suddivise nelle 6 sottosezioni descritte nella tab. 6. Nella categoria "approfondimento" rientrano i programmi come *Annozero*, *Report*, *In mezz'ora*, *Ballarò*, che costituiscono appunto un approfondimento delle notizie principali. "Talk show" sono invece le trasmissioni di argomento più conviviale come *Parla con me* o *Le invasioni barbariche*.

	tokens	% totale
INFORMAZIONE		
telegiornali	229.324	19,99
approfondimento	345.929	30,16
TOT. INFORMAZIONE	*575.253*	*50,15*
INTRATTENIMENTO		
talk show	221.008	19,27
fiction	127.026	11,07
sport	113.181	9,87
spettacolo	110.683	9,65
TOT. INTRATTENIMENTO	*571.898*	*49,86*

Tabella 6 - Sottosezioni della sezione tv

2.10 Film

La sezione comprende la trascrizione integrale dei dialoghi di 66 film italiani prodotti tra il 1995 e il 2011. Le trascrizioni sono state ottenute at-

[2] La ridotta estensione dei dati raccolti per i testi di canzoni (Werner, 2012) ha motivato il loro inserimento nella sezione del parlato monologico anziché una sezione autonoma del PEC.

traverso alcuni siti di condivisione di sottotitoli di film (come *opensubtitles.org*), e successivamente controllate e corrette manualmente.

3 Annotazione

Il PEC è dotato di un'annotazione multilivello, che comprende due fasi distinte: l'annotazione della struttura dei testi e l'annotazione linguistica.

3.1 Annotazione della struttura dei testi

I testi che compongono il PEC sono stati in primo luogo annotati in linguaggio XML, per distinguerli ed etichettarli a livello di genere testuale. Ad un livello ulteriore di dettaglio, ciascun testo è stato etichettato in base alle sue caratteristiche più specifiche: nel parlato, ad esempio, sono annotati i singoli turni di parola e alcune caratteristiche sociolinguistiche dei parlanti (ove possibile, sesso, età e provenienza geografica). E' stato utilizzato un set di tag analogo a quello dello standard della *Text Encoding Initiative* (Burnard e Bauman, 2014); la scelta è stata quella di adottare un tipo di annotazione minimalista, basato su alcune raccomandazioni essenziali (Hardie, 2014).

L'esempio che segue mostra l'annotazione XML di un testo parlato dialogico, prodotto nel corso di un'interazione faccia a faccia da un parlante di 25 anni, di sesso maschile, proveniente dalla Calabria:

```
<text id="427" type="par"
sub="dialogo">
<div type="pari" sub="faf">
<u who="L" sex="m" age="25"
prov="Calabria">
```

3.2 Annotazione linguistica

Il PEC è stato annotato per categoria grammaticale; il pos-tagging (Tamburini, 2007; Attardi e Simi, 2009) è stato effettuato usando *TreeTagger* (Schmid, 1994), con un tagset creato ad hoc[3]; il lessico, pur derivato da quello della distribuzione originale, è stato sensibilmente ampliato, fino a quasi 550.000 entrate. *TreeTagger* è stato addestrato con testi annotati manualmente, appartenenti a tutte le sezioni del corpus: il training set conteneva infatti, in misura uguale, campioni casuali estratti da ciascuna delle 10 sezioni (10.000 parole per sezione, per 100.000 parole totali), per fare in modo che ciascuno dei dieci generi testuali, con le sue peculiarità linguistiche, contribuisse in misura uguale al training del tagger (Jurafsky e Martin 2000; Giesbrecht e Evert, 2009).

La valutazione dell'accuratezza del pos-tagging, effettuata su un test set di oltre 22.000 parole, ottenute in modo bilanciato dalle 10 sezioni del corpus, ha evidenziato un valore del 97,3% (range = 96.6%-97.7%)[4]: la fig. 1 mostra una certa uniformità nei valori delle varie sezioni, con accuratezza leggermente più bassa nei testi parlati.

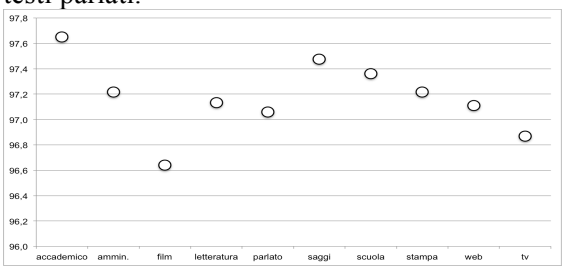

Figura 1 - Accuratezza del pos-tagging nelle 10 sezioni del PEC.

In una fase di "post-tagging", successiva all'annotazione, una serie di errori ricorrenti è stata corretta in modo automatico con l'aiuto di un guesser, basato su espressioni regolari, conformemente a quanto suggerito da Schmid et al. (2007). In tal modo, il pos-tagging ha superato il 98% di accuratezza.

4 Conclusioni

Il PEC rappresenta il primo corpus di riferimento dell'italiano contemporaneo scritto e parlato; nella sua composizione è stata privilegiata la differenziazione dei generi testuali, anche parlati, rispetto all'ampiezza delle dimensioni. Realizzato con risorse limitate e in tempi ristretti, attingendo, ove possibile, a risorse linguistiche già esistenti, il PEC costituisce un compromesso low cost tra creazione di risorse nuove e riuso di risorse esistenti.

L'interrogazione del PEC avviene attraverso l'interfaccia CWB e il *Corpus Query Processor* (Evert e Hardie, 2011), che consente di ricercare parole, sequenze di parole e annotazioni; è prevista la realizzazione di un'interfaccia di rete via *CQPweb* (Hardie, 2012), accessibile al pubblico[5].

[3] Il tagset (http://perugiacorpus.unistrapg.it/tagset.htm) comprende 53 etichette e trae spunto da quello descritto in Baroni et al. (2004).

[4] Sono stati conteggiati sia gli errori di categoria grammaticale che quelli di lemma.

[5] Tale interfaccia consentirà di interrogare il corpus online; non è invece prevista, per motivi di copyright, la disponibilità dei testi che compongono il corpus.

References

Federico Albano Leoni. 2007. Un frammento di storia recente della ricerca (linguistica) italiana. Il corpus CLIPS. *Bollettino d'Italianistica*, IV, (2), pp. 122-130.

Giuseppe Attardi e Maria Simi. 2009. Overview of the EVALITA 2009 Part-of-Speech Tagging Task. *EVALITA 2009*.

Marco Baroni, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston e Marco Mazzoleni. 2004. Introducing the *"la Repubblica"* corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. *Proceedings of LREC 2004*.

Marco Baroni e Silvia Bernardini (eds.). 2006. *Wacky! Working papers on the Web as Corpus*. GEDIT, Bologna.

Marco Baroni e Adam Kilgarriff. 2006. Large Linguistically-Processed Web Corpora for Multiple Languages. *EACL 2006 Proceedings*, 87-90.

Lou Burnard. 2007. *Reference Guide for the British National Corpus (XML Edition)*. Oxford University Computing Services, Oxford.

Lou Burnard e Syd Bauman. 2014. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium, Charlottesville.

Tullio De Mauro, Federico Mancini, Massimo Vedovelli e Miriam Voghera. 1993. *Lessico di frequenza dell'italiano parlato*. EtasLibri, Milano.

Paul Drew e John Heritage (Eds.). 1992. *Talk at Work*. Cambridge University Press, Cambridge.

EAGLES. 1996. *Preliminary recommendations on Corpus Typology* EAG--TCWG--CTYP/P. Version of May, 1996.

Stefan Evert e Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. *Proceedings of the Corpus Linguistics 2011 conference*, University of Birmingham, UK.

Eugenie Giesbrecht e Stefan Evert. 2009. Part-of-speech tagging - a solved task? An evaluation of POS taggers for the Web as corpus. In I. Alegria, I. Leturia, and S. Sharoff (Eds.). *Proceedings of the 5th Web as Corpus Workshop (WAC5),* San Sebastian, Spain.

Paul Guiraud. 1954. *Les Charactères Statistiques du Vocabulaire. Essai de méthodologie*. Presses Universitaires de France, Paris.

Andrew Hardie. 2012. CQPweb - combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17 (3): 380-409.

Andrew Hardie. 2014. Modest XML for Corpora: Not a standard, but a suggestion. *ICAME Journal* 38: 73-103

Marianne Hundt e Geoffrey Leech. 2012. Small is Beautiful – On the Value of Standard Reference Corpora for Observing Recent Grammatical Change. In T. Nevalainen & E. Traugott (Eds). *The Oxford Handbook of the History of English.* Oxford University Press, Oxford, pp. 175-188.

Daniel Jurafsky e James H. Martin. 2000. *Speech and language processing: An introduction to Natural Language Processng, Computational Linguistics and Speech Recognition*, Prentice-Hall, Upper Saddle River, NJ, USA.

Adam Kilgarriff e Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3):333–347.

Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci e Vito Pirrelli. 2014. The PAISÀ Corpus of Italian Web Texts. *Proceedings of the 9th Web as Corpus Workshop (WaC-9),* Association for Computational Linguistics, Gothenburg, Sweden, April 2014. pp. 36-43

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář V, Pavel Rychlý e Vit Suchomel. 2013. The TenTen Corpus Family. *7th International Corpus Linguistics Conference*, Lancaster.

Pascual Pérez-Paredes e Jose M. Alcaraz-Calero. 2007. Developing annotation solutions for online data-driven learning. *EUROCALL 2007* - University of Ulster, 5 - 8 September**.**

Rema Rossini Favretti, Fabio Tamburini e C. De Santis. 2002. A corpus of written Italian: a defined and a dynamic model, in A. Wilson, P. Rayson, T. McEnery (eds.). *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*. Lincom-Europa, Munich.

Helmut Schmid. 1994. *Probabilistic part-of-speech tagging using decision trees*. In Proceedings of the International Conference on New Methods in Language Processing.

Helmut Schmid, Marco Baroni, Eros Zanchetta e Achim Stein. 2007. Il sistema "TreeTagger arricchito". EVALITA 2007. *Intelligenza artificiale*, IV, 2007, 2, pp.22-23.

Stefania Spina. 2005. Il Corpus di Italiano Televisivo (CiT): struttura e annotazione, in Burr, E. (ed.), *Traditione & Innovazione. Il parlato: teoria – corpora – linguistica dei corpora*, Atti del VI Convegno SILFI (28 Giugno - 2 Luglio 2000, Gerhard-Mercator-Universität Duisburg, Germania). Franco Cesati, Firenze, pp. 413-426.

Stefania Spina. 2010. AIWL: una lista di frequenza dell'italiano accademico, in Bolasco S., Chiari I., Giuliano L., *Statistical Analysis of Textual Data*, Proceedings of the 10th JADT Conference (Rome, 9-11 june 2010), Editrice universitaria LED, pp. 1317-1325.

Fabio Tamburini. 2007. Evalita 2007: The Part-of-Speech Tagging Task. *Intelligenza artificiale,* IV, N° 2, pp. 4-7.

Valentin Werner. 2012. Love is all around: a corpus-based study of pop lyrics. *Corpora*, Vol. 7 (1), pp. 19-50

Antonio Zampolli. 1991. Towards reusable linguistic resources. *Proceedings of the fifth conference on European chapter of the Association for Computational Linguistics*.